Seach Products Vendor Analysis

1 Index all content and recognize authentication

Problem Currently there is information on some Web serves that we can't index. For example, anything on the UNIX server that is in a SLACONLY sub-directory information on our intranet server (www-internal).

Requirement Capability of indexing all content, and recognize authentication involved in gaining access to material. Along with this also need to be able to customize the search results display so that no private information is accidentally displayed until authentication is provided.

	Available Out of the Box	Can be done	Not a Feature
Google	Tthe Google Search Appliance supports any HTTP or HTTPS enabled content. Additionally, Basic and NTLM authentication is supported for secured content searching.		
ht://Dig	Yes. Just a line in the config needed to specify what is needed, and what type of content and limits to find and index.		
Inktomi	Inktomi can handle Basic authentication (e.g., .htaccess) quite well – that is how it indexes HyperNews.	It cannot handle https without an additional licensed module.	
Swish-e	Yes. Just config's to set to gather info from servers, and if each server requires a password or not.		
YourAmigo	Yes. (https, Basic and NTLM authentication are supported in July 2003 release).		

Seach Products Vendor Analysis

2 Automated identification of changes

Problem Indexing frequency should be based on frequency of files being updated, so pages that are updated more frequently are indexed more frequently. This is currently done manually based on changes being made.

Requirement Capability of being able to automate the process of identifying areas that are changed frequently and updating the index accordingly (or real-time update; see ID 4).

	Available Out of the Box	Can be done	Not a Feature
Google		Currently, the Google Search Appliance performs a batch-update process taking into account the Last Modified Date of each document. Future versions will feature automatic content crawling based on how often content is updated.	
ht://Dig		Need to setup side jobs to determine new content, and new content can then be merged into main index. A little tricky to setup, but has nice feature of indexing a list of url's, so a list of new content can be fed to the engine as a list.	
Inktomi	Has the ability to schedule collection spidering. In theory, "high change" areas could be scheduled multiple times per day.		
Swish-e		No, but re-indexing can be done fairly rapidly, and searching can continue while re-indexing is going on. Plus with searching multiple indexes, a incremental index can be setup with new changes since the main index. Re-indexing and incremental indexing will need a little work to config and setup as cron jobs.	
YourAmigo	YourAmigo provides an alternative solution to this problem, being of a fundamentally different architecture to the traditional, spider-based engines, for which this so-called “adaptive spidering” is intended to provide a solution. With the YourAmigo agent-based architecture, indexing frequency is less of an issue than with traditional spider-based engines. The agent, co-located with the web server, can scan for changes much more frequently than a spider-based solution because it uses essentially no bandwidth and only sends the changed content.		

Seach Products Vendor Analysis

3 Index content that spans Web servers

Problem When Web content for a given topic spans multiple Web servers there is no easy way to build an index to search just that content.

Requirement Capability of building an index that covers the specific topic, regardless of where it lives in the web architecture and that updates a central index.

	Available Out of the Box	Can be done	Not a Feature
Google	Specific search filters can be created either by URL patterns or by HTML meta tag values.		
ht://Dig	Yes and No.	Would have to multiple index jobs and then merged together at the end. With reasonable limit of 100,000 pages per index, this could limit how far content could indexed.	
Inktomi		Requires building a specialized index – takes admin privs and continued maintenance. Once a collection is created, it does support shared administration at the collection level.	
Swish-e	Yes. Just a set of configs in the spider program to define multiple servers, and each can be config'ed separately.		
YourAmigo	Yes. Can be implemented using either collections or meta data based restrictions. Assignment of documents to collections may be done arbitrarily. Any meta data element may be incorporated into the search criteria.		

4 Efficient indexing

Problem Currently our indexing capabilities are inefficient because 1) the collection is so large, and 2) the indexing doesn't take into consideration frequency of use. In addition, because of the size of our collection, it currently takes a month for the indexer to go through the site. Currently run indexing at night because the indexer depletes user bandwidth (slows down connection to Web pages).

Requirement Capability of efficiently building a comprehensive and real-time index. Indexing to be scheduled on a regular basis, and having the shortest possible time to go through the collection. Want way to indicate when index was last updated. Along with this also need capability of distributed administration function. For example, allow Web authors to indicate when a re-index is necessary (such as when they do a redesign of a site) and what parts of their site should be indexed.

	Available Out of the Box	Can be done	Not a Feature
Google		See answer to #2..	
ht://Dig	Yes, Sort of. Depends what you call eff., it can index about 5,000 pages per hour. But you have complete control of when index is done, you have to setup the cron job to do the index, so it is easy to tell people when the index was created. Since it uses the web server for indexing, it does put load on the server.		
Inktomi	Can index via directory structure, not just via spidering.		
Swish-e	Available of of box: Yes, much more than any other tested. Through the web server it can index 10,000 to 20,000 pages an hour, mostly limited by rate of serving. If the spider is not used, and the information gathering program goes after just web files in the file system, it can index 100,000 files an hour, again limited by file access times. With tuning of file access it perhaps could go faster. We control when indexing is done so it easy to tell people when it was done. Indexing files brings in the ability to create a search index without any load on the web server.		
YourAmigo	Yes. Each YourAmigo agent can run on a separate schedule. Where the agent is installed on the web server machine (the recommended configuration), scanning for changed content is extremely efficient; the web server load is fully controllable and the only bandwidth used is what is required to communicate the changed content to the search server. The last indexed date may be displayed with each result on the search results page. Each agent may be administered separately, allowing the web site owner to have full control over which content is indexed, the indexing schedule, and when to force re-indexing.		

Seach Products Vendor Analysis

5 Index structured, unstructured, and dynamically-created content

Problem Currently can't index structured and dynamically-created content. For example, to search Oracle-based information need to go to special search page (and special page for each other database).

Requirement Be able to effectively index all structured, unstructured, and dynamically-created content.

	Available Out of the Box	Can be done	Not a Feature
Google		The Google Search Appliance supports any data which is available via a HTTP or HTTPS interface. Most databases can be web-enabled using basic programming technologies.	
ht://Dig	Yes, as long as it is served. Notes on work needed: It will grab information from a web server, and how ever the content is created it will index it. But only content server through a web server can be indexed.		
Inktomi			No, and can't index database-driven content.
Swish-e	Yes. It can index through the web server to get exactly what a web browser would see, or call the routines which create the dynamic content and index the output. Also information gather programs can be written to complete replace the spider and index anything: files, e-mail, news groups, databases. This is only limited by ability to re-format things into a text files, and programming ability and time. The distro comes with examples of using a web spider, descending a file system, and MySQL database selects.		
YourAmigo	Yes. YourAmigo supports indexing of dynamic pages backed by SQL databases and implements a unique discovery mechanism which finds the optimum set of URLs covering all published information in the database with minimal redundancy. Primary keys are not mandatory. Multiple fields may be used.		

Seach Products Vendor Analysis

6 Index a wide variety of file formats

Problem Currently indexing of non-html files is inconsistent at best. Some file types have an upper limit, beyond which are not indexed.

Requirement Capability of doing full-text searches on a variety of file formats, such as HTML, XML, Work, PowerPoint, PDF, PostScript, and so on.

	Available Out of the Box	Can be done	Not a Feature
Google	Over 200 file formats are supported by the Google Search Appliance.		
ht://Dig			No. Will only index web pages from a server.
Inktomi	Supports various file formats, but with file size limits and varying degrees of success.		
Swish-e	Yes. Just requires filtering the file format to a text file. The default formats are text, html, and xml for indexing. Examples of filters are provided for gziped files, pdf, and word docs. With a little extra work I have made filters for excel, powerpoint, postscript and all OpenOffice documents. More formats could be defined by programing time to create a filter which translates the format to text or xml. Although use of these filters vastly slows the index creation and uses more memory, which will influence scaling issues.		
YourAmigo	Yes. HTML, XML (fully configurable), MS Word, PowerPoint, PDF, PostScript, Excel, plain text are all supported. External converters may be “plugged in” to convert other formats (e.g. a WordPerfect to HTML converter).		

Seach Products Vendor Analysis

7 Sorted search results

Problem Currently users cannot sort search results very effectively.

Requirement Capability of sorting search results by a variety of methods. Examples are date, group, type, relevancy, and so on. If search categories are defined, then capability of sorting results by category.

	Available Out of the Box	Can be done	Not a Feature
Google	The Google Search Appliance supports sorting of search results by relevancy or date.		
ht://Dig	Yes. Can sort on various data, title, date, and ranking		
Inktomi	Supports sorting of search results by date or by relevancy.		
Swish-e	Yes. Results can be sorted on any metatag information, and arbitrary metatags used in searching can be defined in config's from creating and search the index.		
YourAmigo	Yes. Can sort results by relevancy (default), date, title, URL. Users can also control ranking measures (e.g. turn off popularity, look for words close together etc). If categories are available, can create links to the subset of search results within each relevant category.	Other sorting methods can be added. Typically YourAmigo tries to satisfy customer requests for additional simple features such as this by incorporation into a future release, at no charge. If the feature is required urgently or is very specific to the customer's domain, a quotation can be provided.	

Seach Products Vendor Analysis

8 Assign high ranking in results

Problem Currently have no method to arbitrarily assign higher ranking to specific links. This can result in most valuable links being buried in search results.

Requirement Capability of being able to force a specific result to the top of the search results list. For example, if user enters an "admin" search query, a link to the BIS site should be forced to the top of the search results or show up as a sponsored link.

	Available Out of the Box	Can be done	Not a Feature
Google	The Google Search Appliance has been proven in head-to-head tests to provide highly relevant search results, without manual modification. The appliance does provide a feature called KeyMatch, which allow the administrator to specify editorial results which appear above the standard search results.		
ht://Dig	Yes and No. Ranks by word, not much research was done in how ranking is done. Ranking from information in meta tag and header could increase the ranking of pages.		
Inktomi			No capability.
Swish-e	Yes and no. The ranking of pages is by words searched. You can influence the values of words by where they come in the page, header, title, metatag, body, or other tag (emf or strong). Pages could be brought up in ranking by making metatags more important in the index and putting important words in a metatag for that page.		
YourAmigo	Yes. A ranking adjustment can be assigned to individual URLs.		

Seach Products Vendor Analysis

9 Spelling errors

Problem Currently spelling errors might result in no match or limited matches.

Requirement Capability of recognizing spelling errors and providing alternate search results based on corrected spelling, or provide suggested alternate or similar terms to search on.

	Available Out of the Box	Can be done	Not a Feature
Google	The Google Search Appliance spelling engine begins with the dictionary for google.com and additionally learns the terms from your environment.		
ht://Dig	Yes. There is a fuzzy indexing possibility. This will increase index creation time and index size.		
Inktomi			No capability.
Swish-e	Yes. There are a number of fuzzy indexes which can be created, depending on configs. There are a number of fuzzy algorithms supported. But this may influence index creating time.		
YourAmigo	Yes. Provides a search link with a suggested alternative spelling when no results are found.		

Seach Products Vendor Analysis

10 Search within results

Problem Don't have an easy-to-use search-within-results capability.

Requirement Capability of being able to easily/intuitively search within results.

	Available Out of the Box	Can be done	Not a Feature
Google	Searching within results with the Google Search Appliance is as simple as adding more terms to the search.		
ht://Dig		Not out of the box, but could be done by adding string to the last search.	
Inktomi	Supports search within results, but for untrained users it is not easy to locate.		
Swish-e		Could be added to search interface with some programming work, 1-3 days.	
YourAmigo	Yes. Default is to provide a pre-filled form to which user can add search terms. Alternatives can be easily implemented, by end customer or by request to YourAmigo.		

11 Broken links reporting

Problem Currently don't have good broken link information, which makes correction difficult.

Requirement Capability for an author to query the index for broken links in their Web space, in a report format that is useable and that clearly defines broken link information (such as specific from and to page).

	Available Out of the Box	Can be done	Not a Feature
Google	The Google Search Appliance provides detailed diagnostics about all of the URLs attempted during the crawl, including identification of 404 errors among others.		
ht://Dig			Unknown
Inktomi	Provides a broken links report, but it is not usable because it provides information that is too cryptic to allow research and correction.		
Swish-e		Could be added to spidering program with some programming work, 1-3 days.	
YourAmigo		Planned for later release (nominally Q4 2003). In the interim, other tools (e.g. Linkchecker) are available specifically for broken link checking.	

Seach Products Vendor Analysis

12 Search capability by many kinds of metatags

Problem

Requirement Capability of being able to recognize/search by a large variety of metatags. Examples are content metatags that might define such things as expiration dates, content categorization rules, and personalization rules.

	Available Out of the Box	Can be done	Not a Feature
Google	The Google Search Appliance provides the ability to filter search results based on HTML meta tag values.		
ht://Dig	Yes. But there is a defined list of metatags it will look for.		
Inktomi	Can handle multiple metatagging including all in Dublin Core. Admin is able to weight effects of metatagging. Inktomi help gives illustrations of searching via metatags.		
Swish-e	Yes. Arbitrary meta tags can be defined in configs, to limit search, or search on themselves, or sort on with in a search. Either search or sort on meta tag name or value.		
YourAmigo	Yes. Any meta tag may be added to the list for indexing, and hence searching. Searching can be done either via the query language or by HTML menu options (requires a few lines of ASP/PHP/Perl etc in a search wrapper script).		

Seach Products Vendor Analysis

13 Advanced search logic

Problem Currently requires training to support users in being able to use advanced search logic.

Requirement Query must support advanced search logic, and logic must be relatively straightforward and easy to use. Examples are search by string (name or phrase separated by spaces) or list of keywords (words separated by commas).

	Available Out of the Box	Can be done	Not a Feature
Google	Google has found that >99% of all users simply perform basic search. Google provides excellent search in the default search box, in addition to offering advanced search options for the advanced search user.		
ht://Dig	Yes. There are help docs that come with the distro that people can read for more help.		
Inktomi	Advanced search page is hard to find and requires training to be usable.		
Swish-e	Yes. Includes logic with search string or within meta tags used in searching. Help files comes with distro, but may need a little re-writing for specific uses, or arbitrary meta tags.		
YourAmigo	Yes. Query language uses common operators, e.g. AND/+/OR/|/NOT/-, quotes for phrase search); similar to Google in many ways. Advanced search page is provided. When a user enters a query on the advanced search page, a separate results page appears which shows the query language equivalent of the same query, which effectively provides interactive training.		

Seach Products Vendor Analysis

14 Identify indexing of inappropriate information

Problem Currently we can't identify what part of our Web is open to spidering/indexing by external search engines, which can result in incomplete or inappropriate search results of SLAC-specific material in other search engines.

Requirement Capability of determining if a Web page or site is open to external spidering. We could then hide the information behind SLAC-only, or add it to the robots.txt file, which would eliminate spidering by compliant index tools, or could add appropriate meta tags to reduce the chance the information would be indexed.

	Available Out of the Box	Can be done	Not a Feature
Google	The Google Search Appliance performs spidering similar to other web spiders on the Internet. The appliance can be used as a tool to determine where your public content is located, as well as performing searches for inappropriate material.		
ht://Dig	Yes. Can read robots.txt files, and can be limited by url.		
Inktomi			No capability
Swish-e	Yes. Can be limited by robots.txt, or config's in the spider programs, or arbitrary programs can be made to limit search in any way.		
YourAmigo	Yes. Admin GUI provides tree view of URLs that each agent can access. Unlike spider-based engines, unlinked content (which is sometimes simply material that is overlooked, sometimes benign but poorly managed content, but in some cases can also be a security issue) is also visible in the URL tree. (In July 2003 release) Content which requires authorization is flagged as such.		